Spring 5-15-2015

Investigating the Role of Wolbachia Endosymbionts in the Expansion of the F Element in Drosophila ananassae

Elizabeth J. Chen
Washington University in St Louis

2015

Washington University in St. Louis

Elizabeth J. Chen

Investigating the Role of *Wolbachia* Endosymbionts in the Expansion of the F Element in *Drosophila ananassae*

Abstract

At 4.2 Mb overall, the *Drosophila melanogaster* Muller F element (dot chromosome) is an unusual autosome; it is broadly heterochromatic, but the distal 1.3 Mb has a gene density and expression pattern similar to other autosomes. More intriguing is the large expansion of the *D. ananassae* F element (~20 Mb). Elucidating the factors that contribute to this expansion could improve our understanding of how heterochromatic domains are maintained and amplified.

Previous analyses show that the lateral gene transfer (LGT) of *Wolbachia* (the most widespread intracellular bacteria in the *Rickettsiales* order) into the *D. ananassae* genome is an important contributor to the expansion of the F element. Because many genes in the *Wolbachia* endosymbiont of *D. ananassae* (*wAna*) have not been characterized, I used multiple bioinformatics programs to compare the genome assemblies of *wAna* with *wMel* and *wRi* to improve the *wAna* gene annotations. Collectively, I assigned classifications for ~30% of the *wAna* genes with unknown functions (i.e. predicted hypothetical proteins). Consistent with previous reports, I also found a high density of Insertion Sequence (IS) transposon remnants within the three *Wolbachia* genomes, particularly in *wAna*. These IS sequences might facilitate the LGT of *wAna* and contribute to the expansion of the *D. ananassae* F element.

Analysis of three improved *D. ananassae* F element scaffolds (~1.4 Mb) showed that 65 out of 415 unclassified repeats identified by RepeatMasker have similarity to *wAna*, suggesting that many of these Unknown repeats might be derived from *Wolbachia*. We also compared the distribution of *wAna* genomic scaffolds within introns and intergenic regions as well as identified genomic regions and proteins in *wAna* that are overrepresented in the *D. ananassae* F element.

Collectively, this study will increase our knowledge of the factors that affect chromatin packaging and the evolutionary impact of endosymbionts on host genomes.

Introduction

Unlike genomic DNA in prokaryotic cells, DNA in the eukaryotic genome is packaged into nucleosome arrays, or chromatin, which impacts gene regulation and other cellular activity. There are two major classes of chromatin: the loosely packaged euchromatic regions (which contain actively transcribed protein-coding genes, and the more compact heterochromatic regions (which are enriched for repeats, and other DNA that usually needs to be "off" or not transcribed). Chromatin structure is also altered by epigenetic post-translational modifications of the histones and other chromosomal proteins, such as the methylation of H3 on lysine 9 (H3K9), and by binding specific chromatin proteins such as heterochromatin protein 1 (HP1a); both of these modifications are involved in transcriptional silencing through the formation and maintenance of heterochromatin. These changes to chromatin structure affect the accessibility for the RNA polymerases that transcribe the DNA, and thus are highly regulated. Chromatin structural changes and gene expression misregulation are common causes of many human diseases, including cancer (Jones and Baylin, 2007).

Although much is still unknown about chromatin packaging and gene regulation, progress is being made through the study of the orthologous regions from multiple species (i.e. comparative genomics). Better and faster DNA sequencing, due to innovations in next-generation sequencing technologies, has substantially lowered the costs of sequencing and made comparing genomes of multiple species, such as those in the *Drosophila* lineage, more feasible. *Drosophila* species are readily available for study and useful due to their environmental and biological diversity. Despite phenotypic differences, most of these species have similar cellular

and genetic properties. Sequencing, assembly, and annotation of *Drosophila* genomes can provide additional insights into genomic and evolutionary processes.

One particular area in which *Drosophila* comparative genomics has been utilized is in exploring the properties of the *Drosophila* autosomes. The chromatin in most autosomes is packaged into two major types introduced above: the heterochromatic DNA, tightly packaged and silenced regions, is usually found at the centromeres and telomeres, while the euchromatic DNA, less tightly packaged regions capable of genetic activity or transcription, is found in the arms. However, the *Drosophila melanogaster* fourth chromosome (also known as the dot chromosome or the Muller F element) is predominantly packaged in a heterochromatic form. The *D. melanogaster* F element is small (only 4.2 Mb overall), leading to a metaphase chromosome that looks like a dot. However, the distal 1.3 Mb of the *D. melanogaster* F element has a gene density comparable to those of the other chromosome arms, despite having a repeat density of ~35%. In addition, while this chromosome exhibits heterochromatic characteristics (e.g., high repeat density, with high levels of HP1a and histone H3K9 methylation), the overall expression levels of F element genes are similar to genes in other autosomes (Riddle et al, 2012). Thus this chromosome provides an unusual opportunity to study gene expression in a heterochromatic domain, an environment usually associated with silencing.

Among the different *Drosophila* species, the *D. ananassae* F element is particularly interesting. While the other *D. ananassae* Muller elements have similar lengths compared to their orthologous Muller elements in *D. melanogaster*, the *D. ananassae* F element is substantially larger. (The distal region of the *D. melanogaster* F element is estimated to be 1.3 Mb, while the *D. ananassae* F element assembly is approximately 20 Mb.) Hence investigation into the factors that contributed to the expansion of the *D. ananassae* F element could improve

our understanding of how heterochromatic domains can become enlarged, and the phenotypic impact of this change.

Preliminary studies suggest that one of the contributors to the increase in the size of the *D. ananassae* F element is due to lateral gene transfer from *Wolbachia,* the most widespread intracellular bacteria in the *Rickettsiales* order, which includes species with parasitic, mutualistic, and commensal relationships with the hosts (Serbus et al, 2008). (Some *Rickettsiales* bacteria are also notable pathogens, several of which cause a variety of human diseases, such as Rocky Mountain spotted fever.) Previous studies have reported that the entire *Wolbachia* genome (~ 1Mb) is integrated into the *D. ananassae* genome via lateral gene transfer. This hypothesis is supported by *in situ* hybridization studies that shows the integration of *Wolbachia* fragments into the *D. ananassae* genome (Hotopp et al*,* 2007). Amongst the *Wolbachia* fragments that are integrated into the *D. ananassae* genome, approximately 2% (28 *Wolbachia* genes) are reported to be actively transcribed (Hotopp et al, 2007). Although it appears that multiple copies of the *Wolbachia* genome are present in the *D. ananassae* genome, the number of *Wolbachia* genes and genomic fragments present in the *D. ananassae*'s F element, and their potential contribution to the expansion of F element has not been fully explored.. In addition, previous studies did not examine if there are any potential biases in the distribution and the types of *Wolbachia* sequences that are integrated into the *D. ananassae* genome.

It is estimated that *Wolbachia* can be found in approximately two-thirds of all insect species, and they have been detected in every insect order (Serbus et al, 2008). This success in infecting host genomes is currently hypothesized to be partially due to efficient transmission through the female germline, the tissue in which they are most prominently found. *Wolbachia* are excluded from the mature sperm, which explains the expected low transmission rates – 2% -

through the male germline (Serbus et al, 2008). Confocal microscope imaging and labeling of *Wolbachia* in *D. melanogaster* oocyte development suggests that the transmission of *Wolbachia* to the host occurs in the germline stem cells of infected females. During stem cell mitosis, the bacteria partitions between the self-renewing stem cell and the differentiating cystoblasts. *Wolbachia* are thus both retained within the germline stem cells and transferred into the differentiating daughter cells, which are converted into *Wolbachia*-infected eggs. This infection mechanism enables *Wolbachia* to maintain itself in the germline, presumably making lateral gene transfer possible (Serbus et al, 2008).

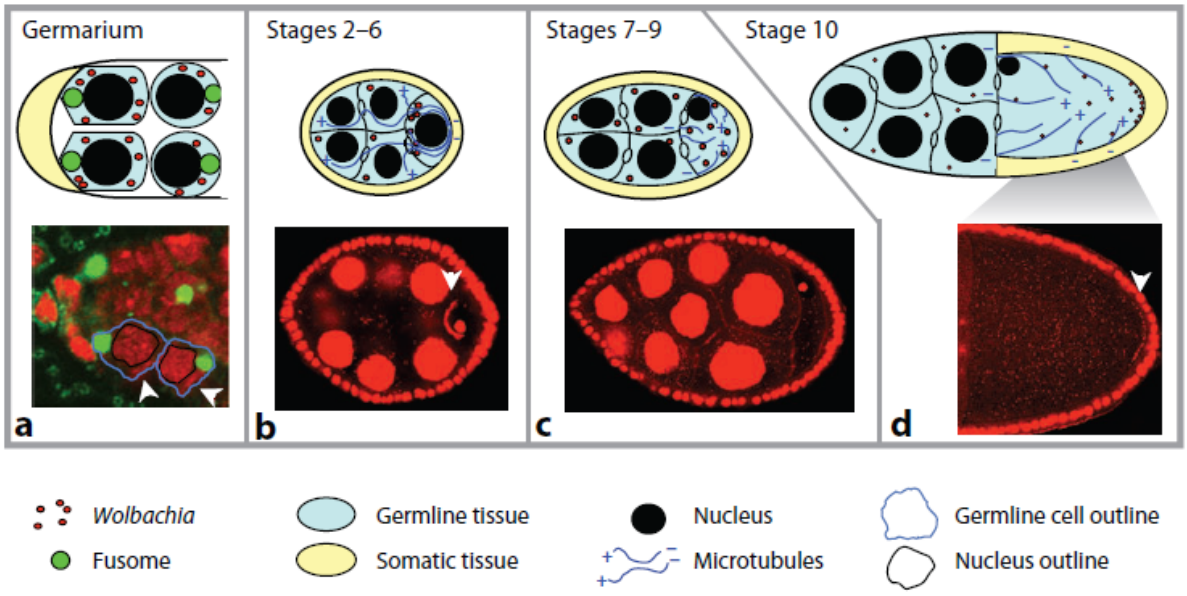


Figure 1: (Serbus et al., 2008) Fluorescence visualization supports the presence of *Wolbachia* during oocyte development. The 16 products of meiosis remain interconnected by ring canals, with 15 nurse cells providing contents for the single oocyte. The fusome is postulated to help form the ring canals.

Genome analysis by Salzberg and colleagues further supports the hypothesis that *Wolbachia* transmissions are primarily through infected females. Their study concluded that eggs or early stage-embryos of the hosts had the greatest amount of *Wolbachia* compared to other infected host cell types (Salzberg et al, 2005). Few, if any, *Wolbachia* are transmitted through the male

germline, as the bacteria are usually eliminated during the final stages of spermatogenesis. However, infection of the sperm has been suggested to play an integral role in inducing cytoplasmic incompatibility (CI), the most common mechanism used by *Wolbachia* to manipulate the host reproductive systems. CI is a form of conditional male sterility, whereby infected males mating with uninfected females results in high mortality rates. Combined with infected females successfully mating with uninfected males, CI is likely one of the mechanisms that explains the prevalence of *Wolbachia* through the insect populations.

In addition to vertical transmission via reproduction, *Wolbachia* also displays success in lateral gene transfer, or horizontal movement across species boundaries, transferring DNA between itself and the hosts. While the molecular mechanisms of this process is still not fully understood, prior experiments have suggested that perhaps some *Wolbachia* strains can briefly exist outside of host cells but then traverse cell membranes, which could aid their horizontal transmission (Werren et al, 2008). Werren and colleagues further argues that there are extensive lateral movements of *Wolbachia* between different *Drosophila* species, given that the phylogeny of the *Wolbachia* species that infect *Drosophila* differs from the established phylogeny of the *Drosophila* species. In contrast, the phylogeny of *Wolbachia* for nematodes is generally the same as the phylogeny for their host species. Their study further suggests that, *Wolbachia* behaves as either parasitic or mutualistic endosymbionts depending on the host species (Werren et al, 2008).

To better understand the transmission and integration of *Wolbachia* and its potential consequences for the host organism, we can use comparative genomics to examine several *Wolbachia*-infected species of the same genus or family. However, before we can perform the comparative analysis, we need to first establish the phylogenetic relationships among the different *Wolbachia* species. Although *Wolbachia* has been found and is known to infect

multiple species of *Drosophila*, it is unclear whether the *Wolbachia* endosymbiont of *D. melanogaster* (*wMel*) or that of *D. simulans* (*wSim*) is the closest relative to the *Wolbachia* endosymbiont of *D. ananassae* (*wAna*) (Salzberg et al, 2005). Current studies have not established whether *Wolbachia* strains in the different *Drosophila* species are endosymbionts (i.e. organisms that live within the cells of another organism and can be either beneficial or harmful) as opposed to a parasite (i.e. organisms which benefit themselves solely at the expense of the host). However, because the official species name for *wAna* in the NCBI taxonomy database is "*Wolbachia* endosymbiont of *Drosophila ananassae*," we will use this nomenclature in this study. Using the published *wAna* assembly as a reference, I will analyze the genomic regions in the *D. ananassae* F element that show significant sequence similarity to *wAna* genomic regions (*Wolbachia* contigs) and *wAna* protein-coding genes (*Wolbachia* proteins), and analyze their impact on the genomic characteristics of the *D. ananassae* F element.

Although prior comparative annotations of the *wAna* genome have used *wMel* as the reference species, previous analysis by Salzberg and colleagues shows that *wAna* genes are more similar to *wSim* (99.8% identity between their nucleotide sequences) than to *wMel* (only 97.2% identity) (Salzberg et al, 2005). In addition, Salzberg and colleagues also found two large gene clusters in *wMel* that are involved in host-endosymbiont interactions that are missing in the *wSim* and *wAna* assemblies. These observations can help explain the difference in nucleotide similarities, and further support the hypothesis that *wAna* is more similar to *wSim* than *wMel*. Although up to this point we have been discussing *wSim*, it should be noted that we will use *wRi* in our analysis. *wRi* is a strain of *Wolbachia* that infects *D. simulans* collected in Riverside, California. The *wRi* and *wSim* genomes are almost identical but the quality of the *wRi* assembly is much higher than the *wSim* assembly because it has been manually improved (Klasson et al,

2009). Previous studies have also shown that the *wRi* genome shows a higher level of sequence similarity to *wAna* than *wSim* (Hotopp et al, 2005). Hence one of the first steps in my analysis of *wAna* is to determine whether I should use *wMel* or *wRi* as the reference species.

Another unusual feature of *Wolbachia* is the high density of Insertion Sequence (IS) transposon remnants in the *Wolbachia* genome. IS elements are small relative to other transposable elements and they usually only contain regions that code for proteins (e.g., transposase) that are involved in their own mobility (Siguier et al, 2005). The transposase within the IS element is usually identified by one or two open reading frames (OrfAB and OrfA), which can consume nearly the entire length of the IS element. Although IS elements usually represent less than 3% of prokaryote genomes, active and remnants of IS elements account for more than 10% of the *Wolbachia* genome (Cerveau et al, 2011). The high density of IS elements in *Wolbachia* could increase the chance of lateral gene transfer and could play a role in the expansion of the *D. ananassae* F element. Prior studies suggest that IS elements are linked to chromosomal rearrangements in other genomes, and a recent comparison of the *wMel* and *wRi* strains identified 17 out of 35 gene-order breakpoints to be flanked by IS elements (Klasson et al, 2009). Therefore, I will also investigate these IS sequences within the three *Wolbachia* species and three improved *D. ananassae* F element scaffolds (improved2_13034, improved_13034, and improved_13010) to see if their presence could potentially play a role in the expansion of the *D. ananassae* F element.

Materials and Methods

Manual Sequence Improvement

In order to better examine the *wAna* fragments that have been integrated into the *D. ananassae* genome, we needed to ensure that the F element analysis regions have been correctly assembled. The original Comparative Analysis Freeze 1 (CAF1) assembly was produced using the Whole Genome Shotgun (WGS) strategy using three libraries with different insert sizes: subclones, fosmids, and bacteria artificial chromosomes (BACs). Because the *D. ananassae* F element has a repeat density of ~80%, the F element scaffolds in the CAF1 assembly contain many errors and ambiguities. To address these assembly issues, I and other undergraduate students participating in the Genomics Education Partnership (GEP) manually improved these sequences using the Phred/Phrap/Consed software package. The common types of assembly issues that other student finishers and I resolved include Single-subclone Regions (SRs), gaps, High-Quality Discrepancies (HQDs), and Low-Quality Regions (LQRs). To address regions that require additional sequencing data, Thomas Quisenberry and I experimented with different PCR protocols to optimize the PCR products for sequencing. These alternate strategies included using specialized enzyme (for the Hot Start Protocol), varying annealing step temperatures (Temperature Gradient Protocol), and varying primer concentrations (Concentration Gradient).

Creating custom tracks

To analyze the distribution of the *Wolbachia* contigs and proteins, I used the custom track functionality of the UCSC Genome Browser to create different custom tracks for the three improved *D. ananassae* F element scaffolds and for the *wAna* genome assembly. I used the Table Browser to collect all data and then used Excel and Notepad to create the BED, plain text, and GFF files. Following the protocol for constructing custom tracks on the UCSC genome browser web site (http://genome.ucsc.edu/goldenpath/help/customTrack.html), I compiled all the sequences from the *wAna* contig of interest, changing the itemRgb of the five IS sequences for

analysis on Genome Browser to create Figures 10, 11, and 12. I similarly created and colored separate tracks for my manually annotated *Wolbachia* transposase (red), gag and pol proteins (brown), and other protein-coding gene fragments (black) for each of the three analyzed *D. ananassae* scaffolds to create Figures 9, 10, 11, and 12.

Classifying *Wolbachia* Genes

First, using the Elgin Lab mirror of the UCSC Genome Browser, I collected the *Wolbachia* genes from the *Wolbachia* endosymbiont genomes of *D. ananassae* (*wAna*), *D. melanogaster* (*wMel*), and *D. simulans* (*wSim* and *wRi*). As explained in the introduction, while I have collected data for both *wSim* and *wRi*, I will only focus on *wRi* in this study. I used the UCSC Table Browser to create BED (i.e. Browser Extensible Data) records for all of the annotated *Wolbachia* genes (available through the "Annotation Genes" track under the "Genes and Gene Prediction Tracks" section). I then imported the BED file into Excel and then examine their GenBank descriptions (all of the entries that begin with "product=") in order to group the *Wolbachia* genes into gene families. Using PivotTables in MS Excel, I determined the number of genes in each gene family for each of the *Wolbachia* endosymbiont genome. I then used these counts to create the pie charts to compare the number of genes found in each gene family in the three different *Wolbachia* species.

Use of *wRi* instead of *wMel* as the reference genome

As described in the introduction, previous studies disagreed on whether *wMel* or *wRi* is the closest relative to *wAna* (Iturbe-Ormaetxe et al, 2005). To determine the best reference genome that I should use in my comparative analysis of the *wAna* assembly, I first compiled the list of *wAna* genes with the GenBank description "conserved hypothetical proteins." Using the

reference protein databases for *Wolbachia* endosymbiont of *Drosophila melanogaster* (taxid:163164) and *Wolbachia* sp. *wRi* (taxid:66084), I performed BLASTP searches of the hypothetical proteins in *wAna* against the annotated proteins in the *wMel* and *wRi* assemblies. From this data, I used R (http://www.r-project.org/) to create a box plot of the percent identities between the *wAna* and the *wRi* protein-coding genes and between the *wAna* and *wMel* protein-coding genes.

Classification of hypothetical protein-coding genes

To further characterize the *wAna* hypothetical protein-coding genes, I searched for the subset of proteins in the *wAna* genome that contain the labels "hypothetical protein" or "conserved hypothetical protein" and recorded their GenBank IDs in a text file. I then uploaded this list of IDs to NCBI Batch Entrez in order to retrieve all of the corresponding *Wolbachia* protein sequences in FASTA format. Using NCBI BLASTP, I searched each sequence separately against the *wMel* and *wRi* proteins in the Reference Sequence (RefSeq) database (refseq_protein) (with the taxid:163164 and taxid:66084, respectively) using an Expect threshold (or E-value) of 1e-10. Although these BLASTP searches resulted in multiple protein alignments for each *Wolbachia* protein sequence, it did not include annotations of the conserved domains, which was one of my criteria in annotating orthologs. To determine if any of these proteins contain conserved domains, I performed an additional BLASTP search against both RefSeq reference protein databases (taxid:163164 and taxid:66084) by using each protein's NCBI sequence identifiers individually, and recorded the conserved domain matches in an Excel workbook. Other evidence used in the annotation and classification of these hypothetical proteins included the protein matches, the percent identity between the hypothetical protein and the RefSeq protein

record, the total score of the alignment, and the length of the alignment. These BLASTP searches were performed using default parameters with an Expect threshold of 1e-10.

Because prior studies and my own analysis of the *wMel* and *wRi* distribution have concluded that *wAna* is more closely related to *wRi* than *wMel*, I first attempted to annotate these hypothetical proteins using the *wRi* reference proteins database. I then used the *wMel* database to annotate the remaining unclassified hypothetical proteins. For example, if a *wAna* hypothetical protein has no significant matches to the annotated proteins in *wRi*, I then performed an additional BLASTP search against the annotated proteins in *wMel* to try to classify the protein.

Based on the aforementioned evidence, I partitioned the BLASTP matches into three categories: ones with strong evidence supporting the ortholog assignment (e.g. supported by the presence of conserved domains, single high quality match detected by BLASTP), ones which had ambiguous evidence (matches to multiple conserved domains and/or protein coding genes), and ones that remained unclassified (e.g. no putative conserved domains or matches only to other hypothetical proteins).

Identification and Calculation of *Wolbachia* gene families

In order to study the distribution of major gene families in *Wolbachia*, I next examined the genes in *wAna, wRi, and wMel.* In order to identify the major gene families in each species, I assign each gene to a gene family based on their GenBank descriptions. Using these assignments I constructed PivotTables in MS Excel to show the number of genes in each gene family for each of the three species. Then I identified the major gene families in each species (defined as the subset of gene families that account for at least 1% of all the annotated protein-coding genes in that species) and this data is used to create the pie charts in Figure 4.

Having defined the major gene families for each species, I then analyzed frequency of each gene family in the other *Wolbachia* species. Using the three lists of gene families, I then took the collection of major gene families in each species and searched it against the other species. I then collected their counts for each species in order to create Figure 7.

Distribution of *wAna* Insertion Sequences (IS)

I searched for the keyword "IS" followed by the keywords "family" or "transposase" in the GenBank records for all of the *wAna* proteins I had previously collected to create a list of all IS elements in the *wAna* genome and their corresponding contig locations. Among all the *wAna* contigs, both AAGB01000018 and AAGB01000003 have the highest density of IS elements. (I will only focus on one of these contigs, AAGB01000018, in this study.) I collected the GenBank IDs (which included: transposase, partial chaperone protein, signal peptidase, ABC transporters, and various transferase and hypothetical proteins) of all the annotated *wAna* proteins in this scaffold. I then created a text file, in which all transposase entries were labeled red, and added this custom track to the Genome Browser to see the distribution of specific gene families (e.g. IS, transposase) on this scaffold.

Calculating the distribution of genomic sequences aligning to *Wolbachia* genes in the *D. ananassae* F element scaffolds

Using the three annotated scaffolds of the *D. ananassae* F element, I identified all regions that show sequence similarity to *wAna* protein-coding genes (*Wolbachia* proteins). In order to more easily evaluate the distribution of these regions, I converted the alignments to each protein into separate alignment blocks by exporting the alignments in GTF (Gene Transfer Format) using the UCSC Table Browser and then manually filtered the results.

Annotation

Thomas Quisenberry, Kevin Ko, and other GEP students examined and reconciled the gene annotations submitted by GEP faculty and students. Collectively, we manually improved and annotated ~1.4 Mb of the *D. ananassae* F element. The improved regions consist of three scaffolds: improved2_13034 (467 kb), improved_13034 (315 kb), and improved_13010 (597kb). The annotators were able to construct gene models for the twelve *D. ananassae* genes that are found within these three scaffolds, which are then being used in a comparative analysis of F element gene characteristics with their corresponding *D. melanogaster* orthologs (Thomas Quisenberry, Senior Thesis, WU 2015).

Intergenic and intron distribution of *wAna* fragments

Using the evidence tracks on the GEP UCSC Genome Browser, I tabulated the total size of the regions that show sequence similarity to *wAna* contigs within the introns of the most comprehensive isoform of the twelve *D. ananassae* F element genes described above. Using the same strategy, I tabulated the total size of the regions within the intergenic regions that show sequence similarity to *wAna* contigs. The "intergenic" regions are defined as the genomic regions between the coding span of the most comprehensive isoform, and the regions before the first gene and after the last gene in each scaffold.

Overlaps with Unknown repeats; other transposons (e.g. LTR, LINE, DNA transposon) identified by RepeatMasker

In order to calculate the overlaps between *wAna* genomic fragments and Unknown repeats in the three *D. ananassae* F element scaffolds, I used the intersection feature of the UCSC Table Browser to determine the regions of overlap between the regions that show

sequence similarity to *Wolbachia* protein-coding genes and regions that are classified as Unknown repeats by RepeatMasker. I then used a similar strategy to construct custom tracks of the other repeat classes from the RepeatMasker track and then identify the intersections between each custom track with the regions that show similarity to *wAna* protein-coding genes. This information was used to calculate the percentage of *wAna* fragments and protein-coding genes that overlap with the transposons identified by RepeatMasker.

Results

Sequence improvement of three *D. ananassae* F element scaffolds

Confirming the *in-silico* assembly using PacBio Reads

To insure the accuracy of my analysis of the distribution of *wAna* genomic fragments and protein-coding genes in the *D. ananassae* F element scaffolds, I used the manually analyzed and improved genomic contigs (labeled by pink boxes in Figure 2) derived from the *D. ananassae* F elements generated by the GEP (see Methods). For the initial set of *D. ananassae* F element projects, the Genome Institute at Washington University produced restriction digest data from fosmid clones. This restriction digest information enabled GEP students to confirm the correctness of the overall assembly as well as to verify the number of copies of repeats and gap sizes. This strategy was used to confirm the assemblies for two of the improved regions (improved_13010 and improved_13034) producing 913 kb of improved sequences.

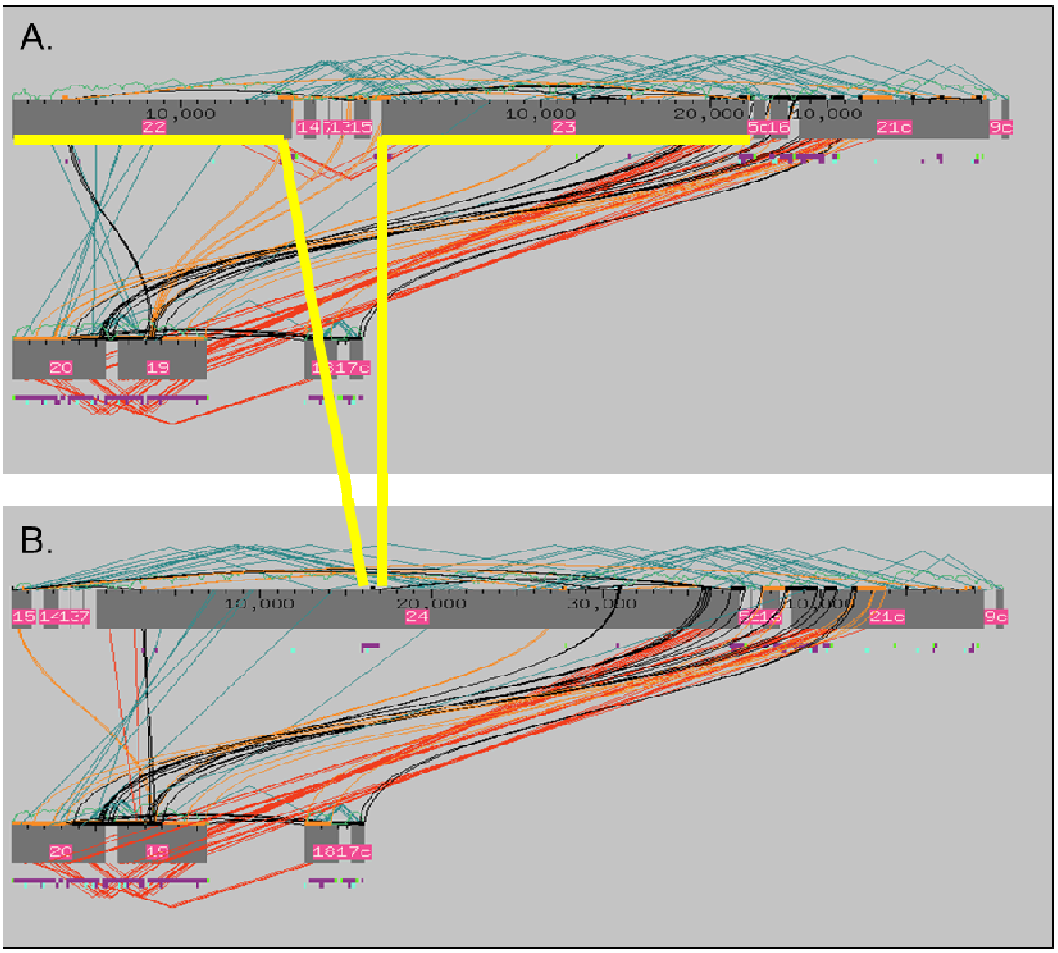


Figure 2: Advantages of manual sequence improvement in resolving assembly issues. A.) The original Assembly View for the finishing project 7278B11. Red lines denote inconsistent mate pairs, orange boxes correspond to direct repeats, and black boxes correspond to inverted repeats. B.) After using PacBio reads to confirm the size of the region around this gap and incorporating additional *D. ananassae* genomic reads from the NCBI Trace Archive, I was able to resolve the gap and inconsistencies between the two larger contigs. The yellow lines between sections A and B denote the region of the assembly that I have resolved.

However, because the *D. ananassae* fosmid clones were no longer available, we used PacBio reads to confirm the integrity of the assembly for the scaffold improved2_13034. Although they have much lower quality than Sanger reads (~80% accuracy), the PacBio data has occasionally helped us resolve some of the assembly issues such as gaps and local misassemblies because of its capacity to generate long sequence reads (Figure 2). Basically, we used the PacBio data in lieu of restriction digests to estimate the size of a gap or misassembly. We could then retrieve additional Sanger reads from the NCBI Trace Archive that showed significant sequence

similarity to the PacBio read in order to fill in the gaps. For cases where no additional Sanger reads were available, the PacBio reads nonetheless provide us with an estimated number of repeats and total gap size. Using this strategy, our Washington University Bio 4342 class (with the help of the professional finishers at the Washington University Genome Institute) improved a 467 kb region of the *D. ananassae* F element. Collectively, all of the GEP students improved a total of ~1.4 Mb of the *D. ananassae* F element closing 26 out of 32 gaps compared to the original *D. ananassae* assembly published by Agencourt (Drosophila 12 Genomes Consortium et al, 2007).

Improvement through PCR Optimization

As is the case for the project shown in Figure 2, some of the projects submitted by GEP students were incomplete (e.g. due to insufficient class time to complete the project) and would benefit from having additional sequencing data. During summer 2014, we attempted to generate the needed data by optimizing the PCR protocol for each case in order to increase the success rate of producing PCR products that would be suitable for subsequent sequencing. Although the standard protocol produced PCR products that resolved a majority of the low quality regions, there were still a few areas that the professional finishers at the WU Genome Institute marked as "DataNeeded", such as Figure 3A.

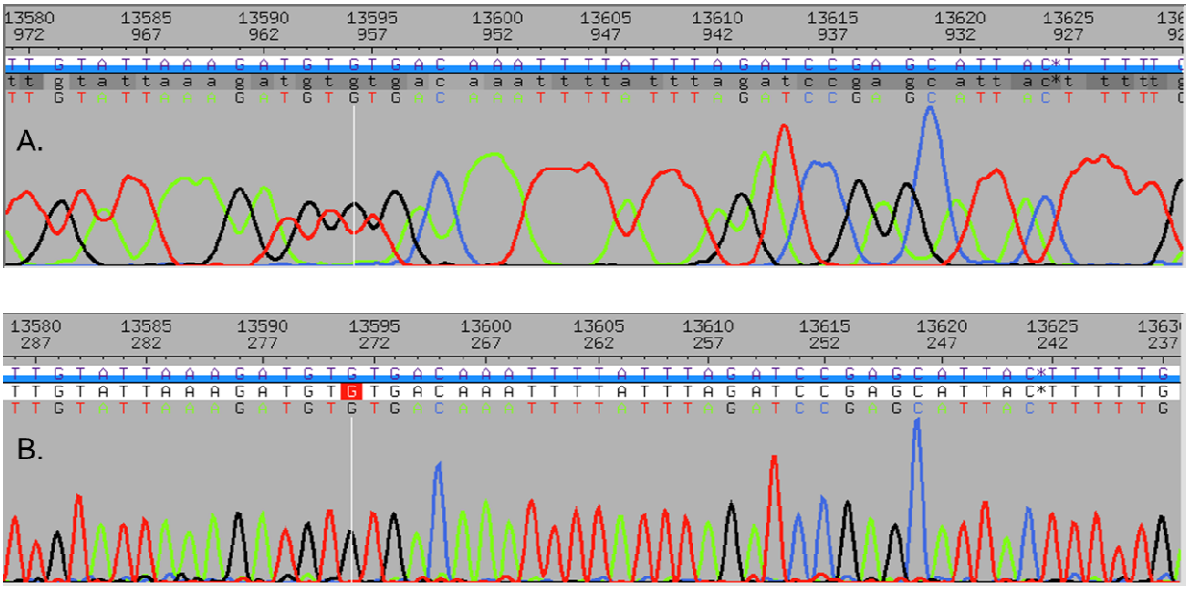


Figure 3: Sequencing using improved PCR products. A. Trace view of the initial sequencing data for a low quality region in the sequence improvement project 5138A08; grey boxes of different shades in the trace view denote low quality bases (i.e. uneven, overlapping peaks). B. The new sequenced PCR product covers the low quality region with high quality data (as denoted by the white uppercase letters and the distinct peaks).

Ultimately, after using three different PCR protocols (see "Methods – Manual Sequence Improvement" for details) on all the genomic regions that require additional data, we were able to generate higher quality traces that improved the quality of the consensus for three out of twelve low quality regions (see Figure 3 for an example). Although we were unable to resolve all the problem areas, Figure 2B and Figure 3B nonetheless demonstrate that we have made substantial improvement to the overall quality of the F element assembly.

Distribution of *Wolbachia* genes and their gene families

Using the data from the Elgin Lab mirror site of the UCSC Genome Browser, I tabulated and classified the genes found in three *Wolbachia* species (*wAna, wMel,* and *wRi*) in order to see if there are any differences in the number and composition of *Wolbachia* genes among the three species. My analysis shows that the *wAna* assembly has more annotated protein-coding genes (1804 genes) than the other species (1169 genes in *wRi* and 1195 genes in *wMel*). In all three

Wolbachia species, the most frequent descriptions of the *Wolbachia* genes were "hypothetical proteins", "transposase" and "ankyrin repeat domain protein". In Figure 4, "Other" corresponds to the other gene families that had GenBank gene description counts lower than 1% in their *Wolbachia* endosymbiont genome assembly (see Materials and methods for details).

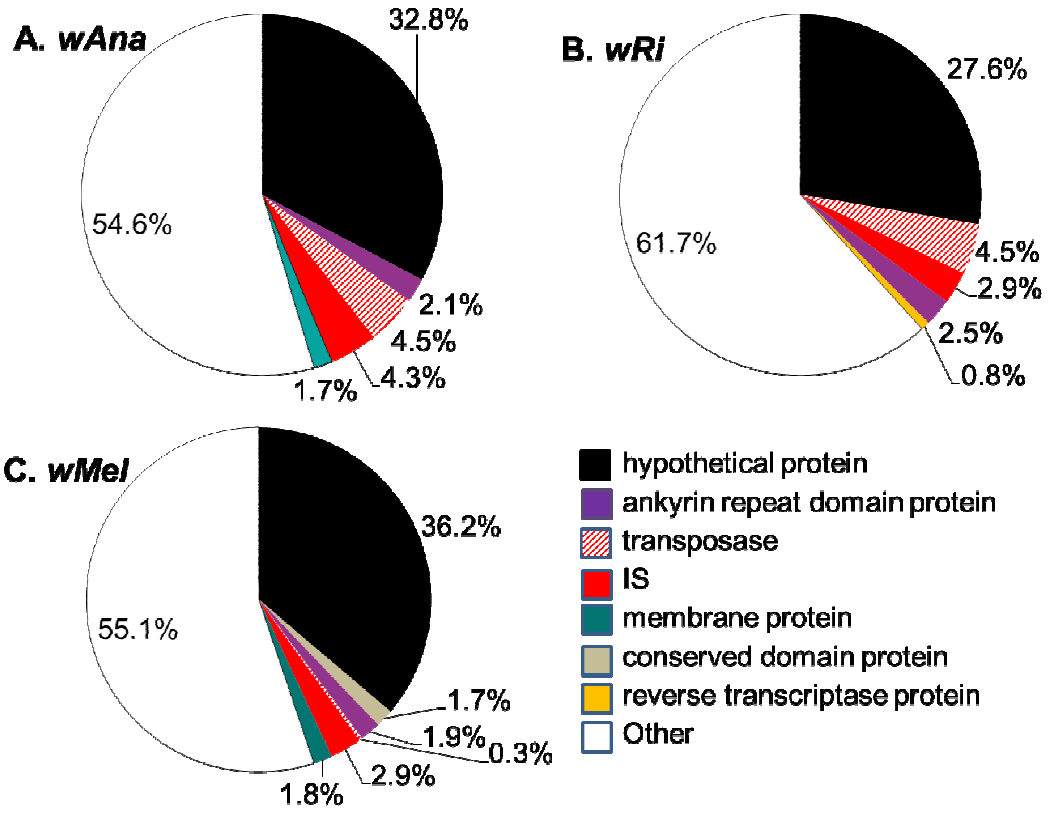


Figure 4: Distribution of Wolbachia protein-coding genes in *wAna, wMel,* and *wRi.* A.) Composition of the 1804 *wAna* protein-coding genes B.) Composition of the 1169 total *wRi* protein-coding genes. C.) Composition of the 1195 total *wMel* protein-coding genes. Not only is there a substantial increase in the number of protein-coding genes in *wAna* compared to *wRi* and *wMel*, there is also an increase in the percentage of IS elements (4.3% in *wAna*, 2.9% in *wRi*, and 2.9% in *wMel*), and transposase (4.5% in *wAna*, 4.5% in *wRi* and 0.3% in *wMel*).

The high percentage of transposase genes and IS elements in *wAna* compared to *wMel* could increase the probability of horizontal gene transfer, since more of the transposons might be active in the *wAna* genome. If so, there might also be a similar bias in the *Wolbachia* protein-coding genes that have been integrated into the *D. ananassae* F element scaffolds.

Initial analysis of *Wolbachia* gene fragments in the *D. ananassae* F element scaffolds

Preliminary investigation of the three improved *D. ananassae* F element scaffolds (improved2_13034, improved_13034, and improved_13010) suggests that these genome assemblies contain multiple genomic regions that have sequence similarity to *Wolbachia* (*wAna*) contigs. Many of these regions also show sequence similarity to *wAna* protein-coding genes in *Wolbachia*. However, as suggested by the previous analysis of the major gene families in the three *Wolbachia* endosymbiont genomes, analysis of the *D. ananassae* F element genome assembly shows that most of the regions of the *D. ananassae* F element with similarity to *Wolbachia* protein-coding genes are annotated as hypothetical proteins (Table 1).

Count	GI ID Numbers	Description
27	58533431	conserved hypothetical protein
24	58533655	conserved hypothetical protein
6	58534611	SD27140p
5	58533971	SD27140p
4	58533425	pol protein, partial
4	58534532	reverse transcriptase polyprotein, partial
4	58534659	hypothetical protein WwAna1191
4	58534383	transposase, degenerate
...		
117		Total Wolbachia Fragments

Table 1: Distribution of *Wolbachia* protein coding genes found on the *D. ananassae* F element scaffold improved2_13034. Of the 117 matches to *Wolbachia* protein-coding genes, 51 of them are annotated as fragments of conserved hypothetical proteins. The "Description" column corresponds to the description in the GenBank record. [Note that some of the *Wolbachia* gene records have the same description but they have a different GenInfo Identifier (GI ID), indicating that they correspond to different protein-coding genes in the *wAna* assembly. For instance, the table shows that there are 27 copies of the gene with the GI number 58533431 and 24 copies of the gene with the GI number 58533655. The GenBank description for both genes is "conserved hypothetical proteins."]

These hypothetical proteins gene annotations were from the original *wAna* genome assembly published in 2005 (Salzberg et al, 2005). Because multiple *Wolbachia* genomes have subsequently been sequenced and annotated (e.g., Klasson et al, 2009; Siozios et al, 2013), I

decided to analyze the *wAna* genes described as "hypothetical proteins" to see if I could improve the current *wAna* distribution analysis by classifying additional *wAna* genes. By doing so, I sought to improve the understanding of the distribution of *Wolbachia* fragments in the *D. ananassae* F element.

Using *wRi* instead of *wMel* as the reference

Before I could classify the hypothetical proteins in *wAna*, I need to first determine whether I should use *wMel* or *wRi* as the reference species. Using the 333 conserved hypothetical proteins from the GenBank record of the *wAna* assembly, I used BLASTP to find regions of similarity between these *wAna* proteins and proteins in the *wMel* and *wRi* reference protein databases (Figure 5A). While the alignments between *wAna* and *wRi* proteins have higher percent identity (median 97.63%) than the alignments between *wAna* and *wMel* proteins (median 81.16%), it should be noted that there are many more protein alignments between *wMel* and *wAna* (560 alignments) than between *wRi* and *wAna* (164 alignments). However, there are more unique matches to *wRi* proteins than *wMel* proteins (42 duplicates and 121 unique matches in *wRi* versus 406 duplicates and 154 unique matches in *wMel*). As a control, I compared the *wAna* ankyrin genes to their orthologs in *wMel* and *wRi*. However, the resulting box plots show that these proteins only exhibit weak sequence similarity among *wMel*, wRi, and *wAna* (~40%) (Figure 5B). Consequently, I also aligned the subset of *wAna* hypothetical proteins that have only a single match in both *wMel* and *wRi.* Of the 333 *wAna* conserved hypothetical proteins analyzed in this study that had a significant alignment to proteins in either *wRi* (121 uniquely identified genes) or *wMel* (154 such cases), 42 of the uniquely identified genes aligned to both *wMel* and *wRi*.

Analyses using this subset of 42 protein-coding genes with clear ortholog assignments in both *wMel* and *wRi* (i.e. paired conserved hypothetical proteins) did not change our original conclusion (Figure 5C). Although the difference in percent identity is slightly lower than the original analysis with all hypothetical proteins, the protein alignments between *wAna* and *wRi* still have substantially higher percent identity (median 95.51%) than the protein alignments between *wAna* and *wMel* (median 78.28%). The results suggest that the conserved hypothetical proteins in *wAna* are more closely related to *wRi* than to *wMel*. While the alignments to ankyrin produce a different conclusion, those alignments have very low percent identity (~40%); the percent identity is below the target frequencies of the BLOSUM62 matrix used in my BLASTP searches. Hence the alignments to the ankyrin proteins are less reliable than the alignments to the paired conserved hypothetical proteins. Based on these alignment results, I decided to use *wRi* instead of *wMel* as my primary reference in the comparative analysis.

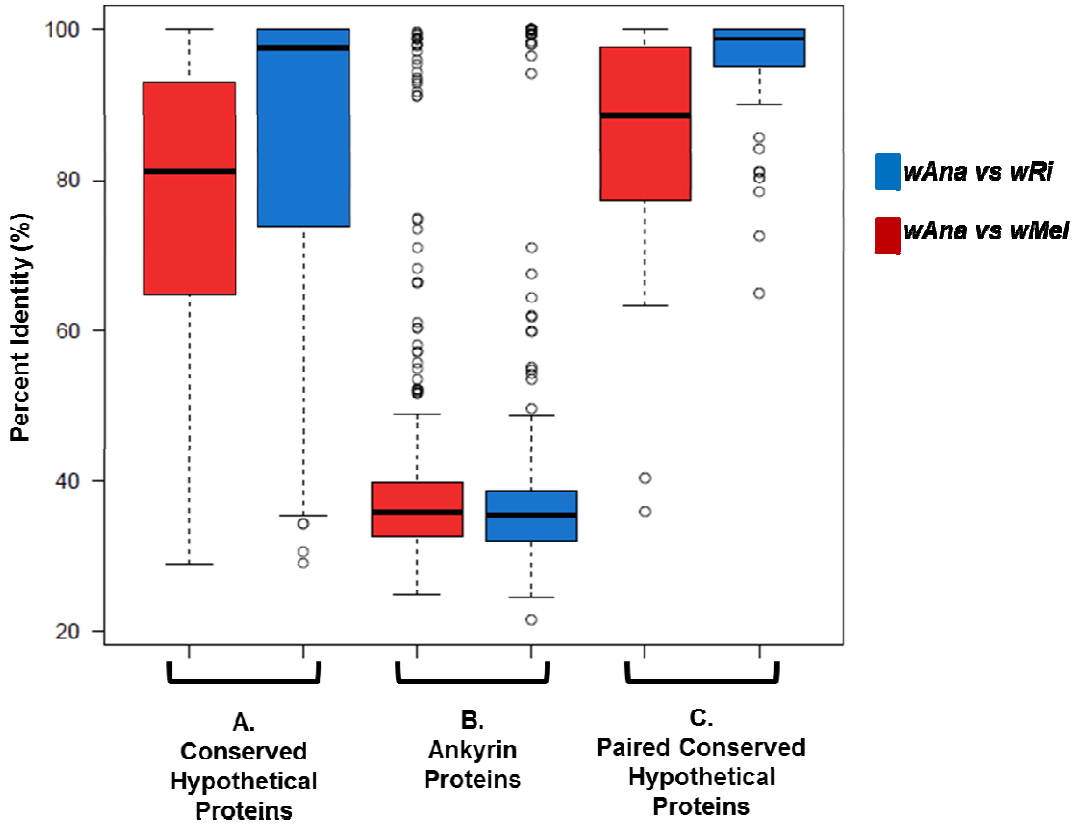
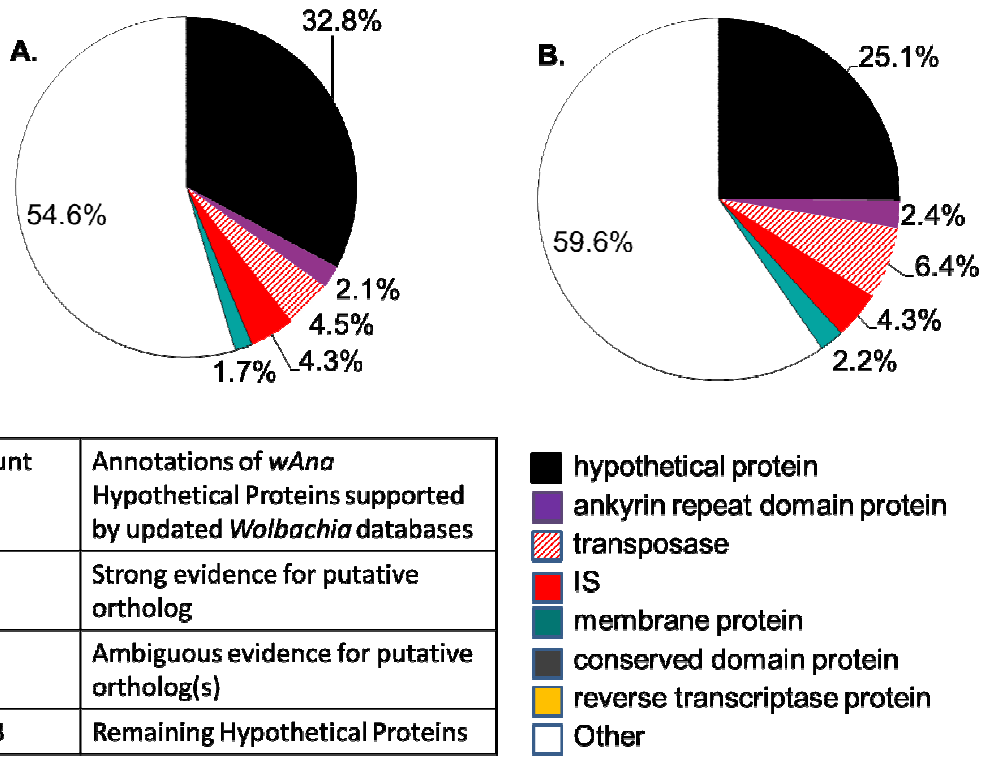


Figure 5: Box Plot of percent identity of *wMel* and *wRi* proteins that aligned to *wAna* proteins as determined by BLASTP. A.) *wAna* Conserved Hypothetical Proteins B.) *wAna* ankyrin proteins C.) *wAna* Subset of Conserved Hypothetical Proteins that aligned to both *wMel* and *wAna*.

Classification of Conserved Hypothetical Proteins

Because the expect values range from 0.0 to the E-value threshold of 1e-10, and the percentage identity values range from ~30% to 100% for the supported orthologous genes, I decided not to rely exclusively on these metrics when I annotated the conserved hypothetical proteins in *wAna*. Instead, I made an ortholog assignment only if it was supported by the BLASTP alignment and the protein is similar in at least one putative conserved domain. This strategy enabled me to classify proteins with low percentage identities (e.g. ankyrin). Using this strategy, I first tried to classify the *wAna* hypothetical proteins based on matches to the *wRi* orthologs. If no matches were found, I then tried to classify the protein based on similarity to the *wMel* orthologs. Collectively, my annotations reduced the number of hypothetical proteins from

32.8% to 25.1% and increased the number of classified transposase proteins from 4.5% to 6.4%. The pie chart in Figure 6 shows the difference between the original *wAna* GenBank annotations and my improved annotations, which shows an increased number of putative transposase, ankyrin, and a few additional membrane proteins in the *wAna* genome assembly.



Count	Annotations of *wAna* Hypothetical Proteins supported by updated *Wolbachia* databases
72	Strong evidence for putative ortholog
67	Ambiguous evidence for putative ortholog(s)
453	Remaining Hypothetical Proteins

Figure 6: Classification of Conserved Hypothetical Proteins. A. Distribution of *Wolbachia* endosymbiont of *D. ananassae* genes before analysis. B. The distribution of *wAna* genes after my classification efforts. Although there are still hypothetical proteins that remained unclassified, most of the hypothetical genes that I have classified are annotated as transposases.

Distribution of *Wolbachia* gene families

To develop a better understanding of the gene distribution within the different *Wolbachia* species and to look for any bias toward specific gene families, I compiled a count of *wAna* genes which had the most frequent gene descriptions (descriptions whose count >1% of the total distribution of gene descriptions of at least one of the *Wolbachia* species), excluding hypothetical

proteins. Although there does not seem to be one consistent trend that applies to all three species, my analysis shows that there are substantially more transposase genes and insertion sequences (IS) in the *wAna* assembly compared to the other species (*wAna* – 82 and 77, *wRi* – 53 and 34, *wMel* – 4 and 35), while the counts for the other gene families did not have as large a difference (Figure 7).

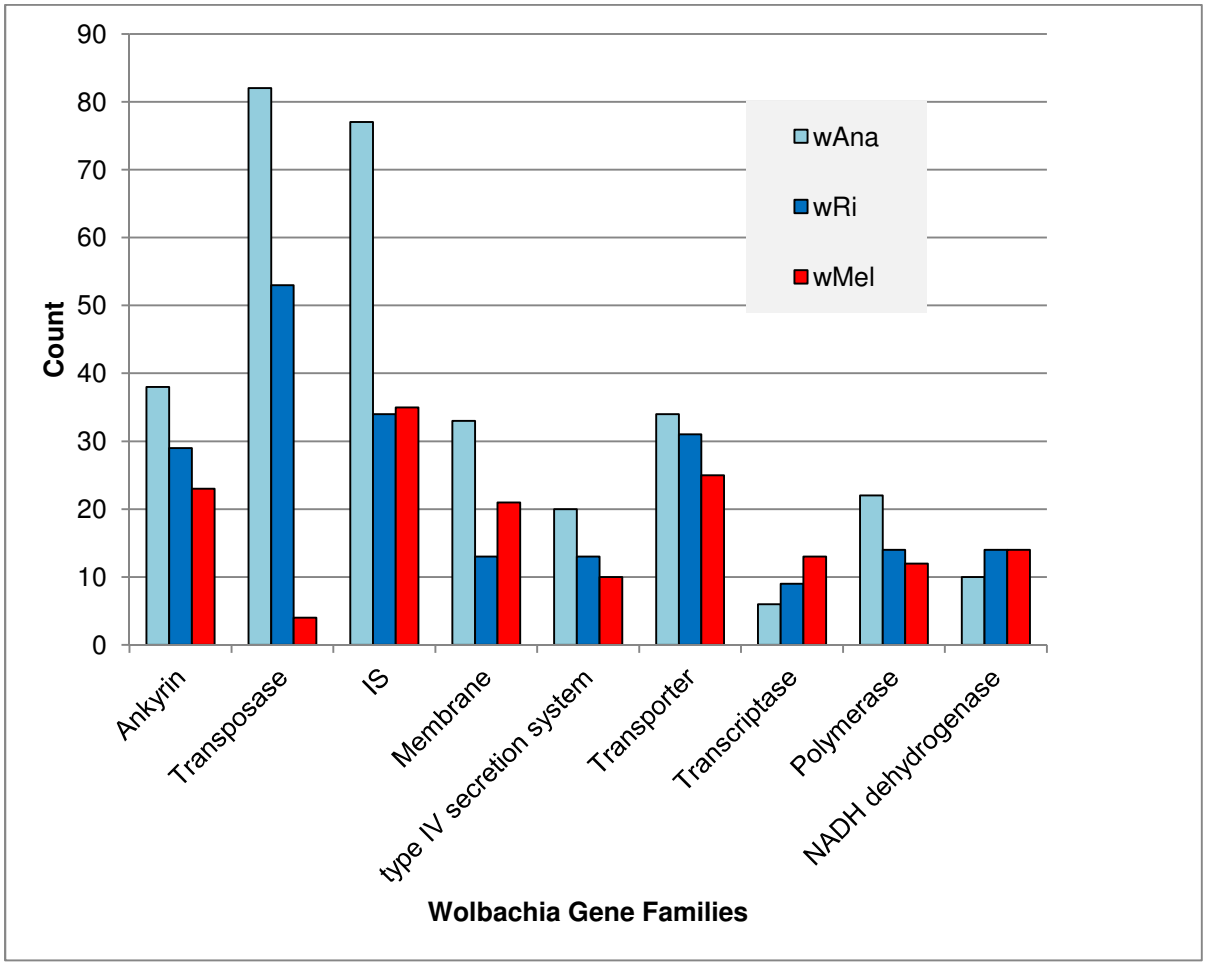


Figure 7: Frequency of a subset of *Wolbachia* gene families that appear at least 1% of the time across *wAna*, *wRi*, and *wMel*. The frequency analysis shows that *wAna* is highly enriched in genes encoding *transposase* and *Insertion Sequences* (IS) compared to the other *Wolbachia* species.

Distribution of *Wolbachia* Insertion Sequences (IS) in *wAna*

To further classify the Insertion Sequences (IS), I searched for GenBank gene descriptions and annotation notes that contain the keywords "transposase" and "IS". Following

the nomenclature for IS elements, the "IS" keyword must be followed either by a number, a number and the keyword “family”, or the “First letter of the genera” and the first two letters of a specific bacterial species (e.g. *ISSod1*). I also classified genes as IS if their gene description included “OrfA”, which is a regulatory protein, and “OrfAB”, which codes the transposase within the IS element (Chandler & Mahillon, 2002). To see if there is a bias in the distribution of the increased number of Insertion Sequences in *wAna*, I then looked at the contigs in which the IS elements are located. Of the IS currently annotated, the majority are interspersed through the multiple *wAna* contigs. However, five of the forty-five *wAna* contigs have more than four identified IS elements. An example of clustering IS-associated elements can be seen in contig AAGB01000018 (Figure 8). In this contig, the IS-associated transposase genes form two clusters, one between 6.5kb to 7.5kb, and one at 9kb to 10kb. Interestingly, there is also a sharp increase in the *D. ananassae* genome coverage within the 9kb-10kb IS cluster, which may be due to inappropriate incorporation of sequencing reads from elsewhere in the *D. ananassae* genome into the *wAna* genome. This observation suggests that this cluster of transposase might be derived from a transposase in the *D. ananassae* genome rather than the *wAna* genome. However, the first two transposase have been noted to contain OrfA and OrfB, which are confirmed *Wolbachia* genes.

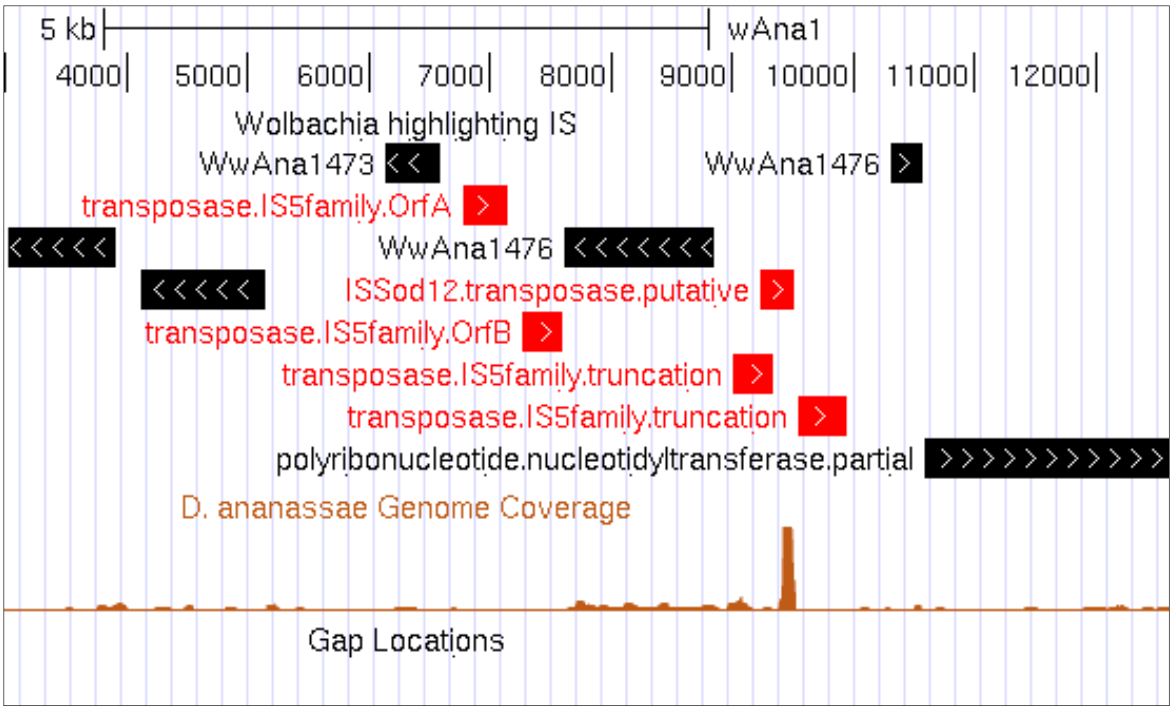


Figure 8: Genome Browser view of a region in the *wAna* assembly with a high density of IS transposons. Many of the genes encoding transposase proteins (red boxes) cluster in two adjacent regions (e.g. at ~6.8-10kb of the scaffold).

Comparison of 470kb of the *D. melanogaster* and *D. ananassae* F elements

Using the higher quality annotation of the *wAna* assembly, I then began my analysis of the distribution of *Wolbachia* fragments in the *D. ananassae* F element by searching for the *Wolbachia* fragments that had been annotated as "conserved hypothetical protein." Among all the *Wolbachia* protein-coding genes found on the three *D. ananassae* F element scaffolds, 13 of them are identified as "conserved hypothetical proteins." However, 12 of these proteins only match to hypothetical proteins in the other *Wolbachia* species. The remaining protein is classified as a DNA topoisomerase 2-like protein in the other *Wolbachia* species. Interestingly, I also found multiple copies of the other conserved hypothetical proteins in the three *D. ananassae* F element scaffolds. However, given the small sample size, I could not identify any patterns in regard to the distribution of these hypothetical proteins on the *D. ananassae* F element.

Next, in order to compare the genomic landscape of the *D. melanogaster* and *D. ananassae* F elements, I examined the first 467,128 bp of the *D. melanogaster* F element compared with the improved *D. ananassae* F element scaffold improved2_13034 using the GEP UCSC Genome Browser. This analysis shows that the *D. ananassae* F element has a much lower gene density (five genes) than the *D. melanogaster* F element (27 genes). Of the five annotated genes (*Crk, Arf102F, Mitf, Zip102B,* and *lgs*) in the *D. ananassae* scaffold improved2_13034, only three of the genes (*Crk*, *Zip102B*, and *lgs*) are found in the first 470 kb of the *D. melanogaster* F element (Figure 9). To further investigate this substantial decrease in gene density, I included a custom track of the regions of the *D. ananassae* F element that shows similarity to *wAna* contigs, *wAna* protein-coding genes, and transposons identified by RepeatMasker. The high density of *wAna* contigs and protein-coding genes within this improved scaffold of the *D. ananassae* F element supports the hypothesis that the decreased gene density is likely due to the integration of *wAna* into the *D. ananassae* F element (27,121bp of DNA in this 467,128 bp scaffold is *Wolbachia*).

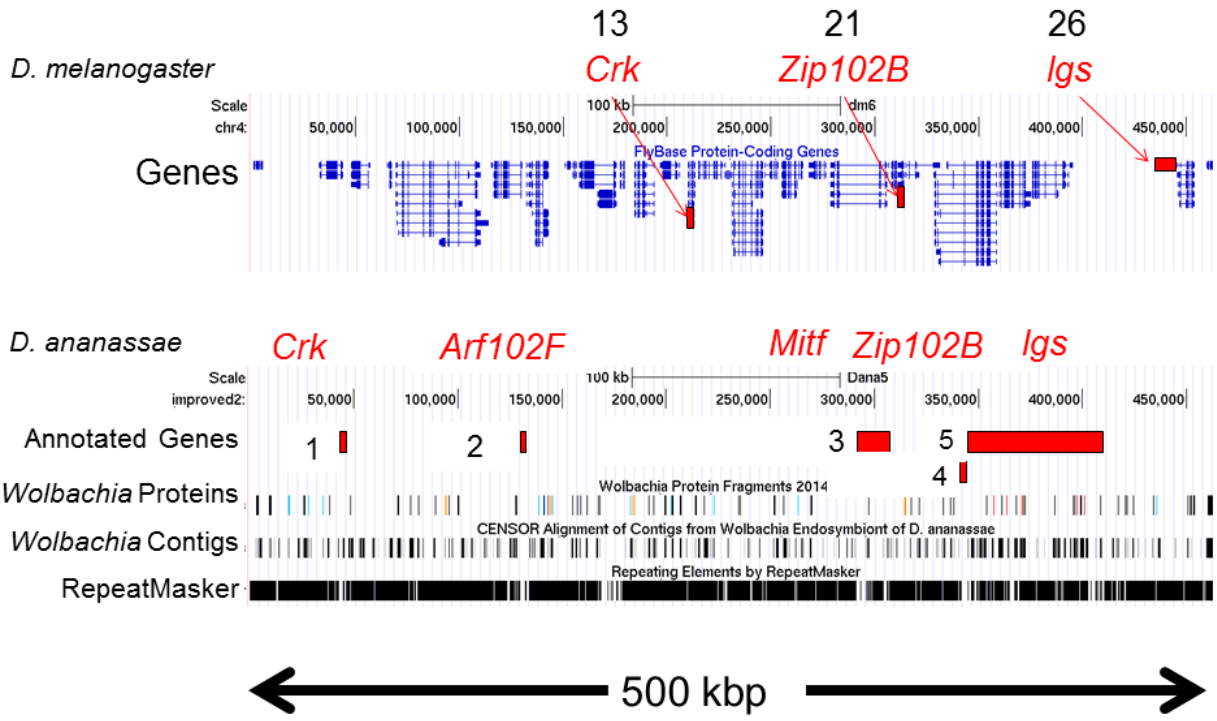


Figure 9: Comparison of the distribution and number of annotated genes in 1-467,128 bp on the *D. melanogaster* F element to that in 1-467,128 bp on the *D. ananassae* F element 13034 scaffold. The number next to each red box corresponds to the gene number relative to the start of the region.

Comparing *Wolbachia* contigs within intron regions versus within intergenic regions

Of the three *D. ananassae* F element scaffolds used in this analysis, only improved2_13034 and improved_13034 have *Wolbachia* protein-coding genes that code for transposase (Figure 10A and 10B). Interestingly, gag and pol proteins are only found in the *wAna* genes (6 and 9 gag and pol proteins in *wAna* but 0 in both *wRi* and *wMel*). However, it is unknown what role these genes might play in expanding the size of the *D. ananassae* F element as the distributions of gag and pol vary between the three scaffolds.

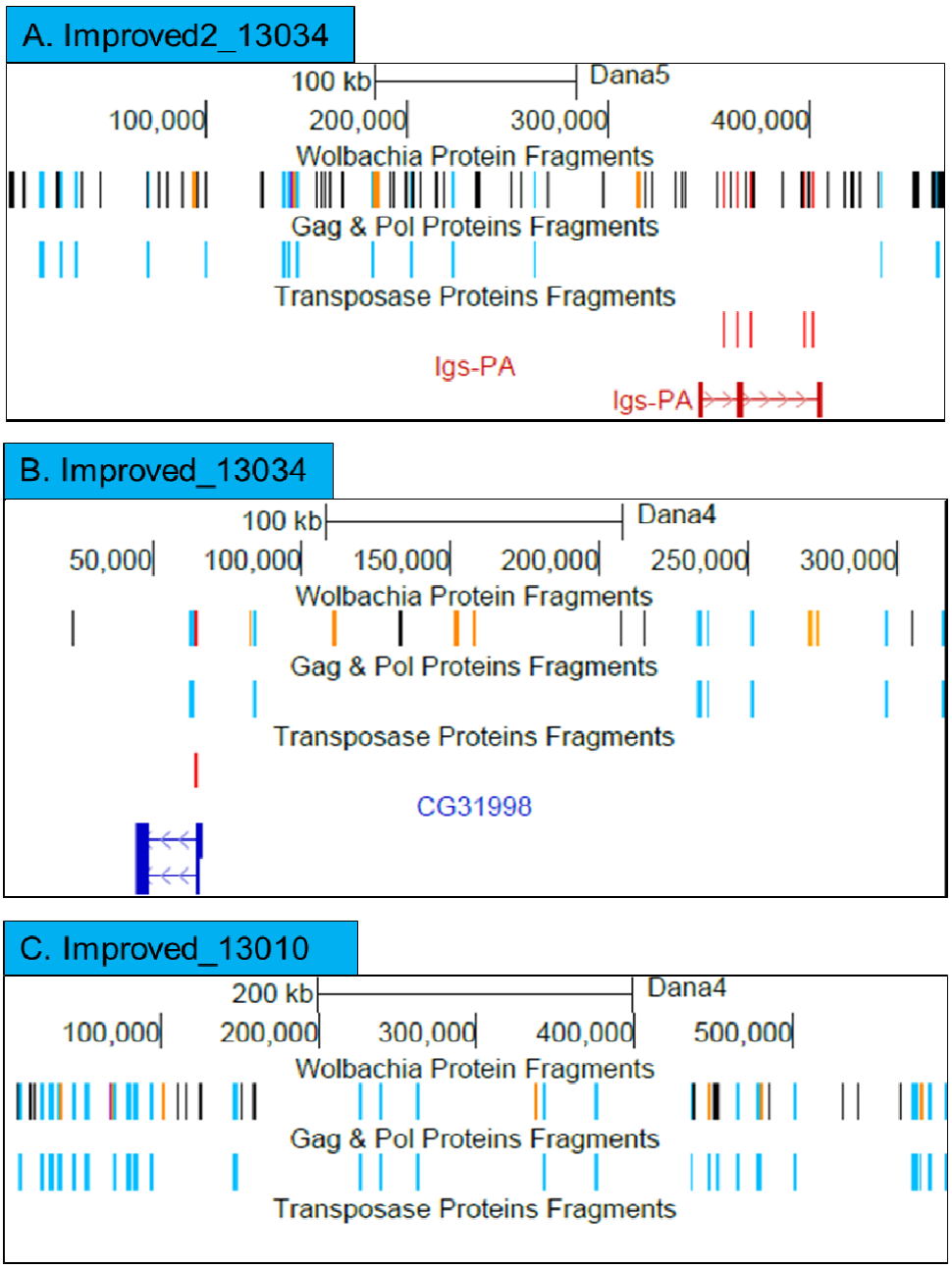


Figure 10: Genome Browser view of *Wolbachia* contigs and proteins-coding genes that overlap with Unknown repeats identified by RepeatMasker. 58 Wolbachia protein-coding gene fragments out of 103 overlap with Unknown repeats.; All transposase - red, all gag, pol, and gag-pol - light blue, reverse transcriptase - orange, while ankyrin - dark violet, and other proteins - black.

To calculate the density of *Wolbachia* contigs within the intergenic regions of the three *D. ananassae* F element scaffolds, I calculated the cumulative size of the regions with similarity to *Wolbachia* contigs and then divide it by the total size of the intergenic regions. I used the same approach to calculate the *Wolbachia* density within introns (i.e. divide the cumulative size of

Wolbachia contigs in introns by the total intron size). I find that there are a higher percentage of *Wolbachia* fragments within introns (27.6–34.5%) than within the intergenic regions (23.5-24.6%), suggesting a potential bias towards insertion into genes (Table 2).

F element scaffolds	% *Wolbachia* within Intergenic regions	Total Intergenic size	% *Wolbachia* within Intron regions	Total intron size
Improved2_13034	24.58%	390,102	34.47%	63,883
Improved_13034*	23.88%	255,228	27.56%	50,111
Improved_13010	23.47%	504,195	33.89%	78,037

Table 2: *Wolbachia* distribution in the *D. ananassae* F element intergenic and intron regions. *Improved_13034 contained an annotated ankyrin gene. I omitted the *Wolbachia* regions that overlap with the *Ankyrin* gene from my analysis because the *D. ananassae Ankyrin* gene will show significant matches to the *Wolbachia ankyrin* gene because of the conserved domains that are found in both genes.

Five transposase protein-coding gene fragments were found in the scaffolds improved_13034 and improved2_13034. However, the transposase fragments in improved_13034 are clustered close enough together that it is possible that they could be fragments of the same transposase gene, resulting from a single integration events (Figure 11B). This hypothesis is supported by the fact that all five matches within this cluster have the same GenBank identifiers. On the other hand, even though all the transposase fragments are clustered within the *lgs* gene, some of these transposase fragments are located in different introns (Figure 11A). Furthermore, of the five transposase fragments, only two fragments have the same GenBank identifier. Hence there are likely multiple transposase insertions into the *lgs* gene.

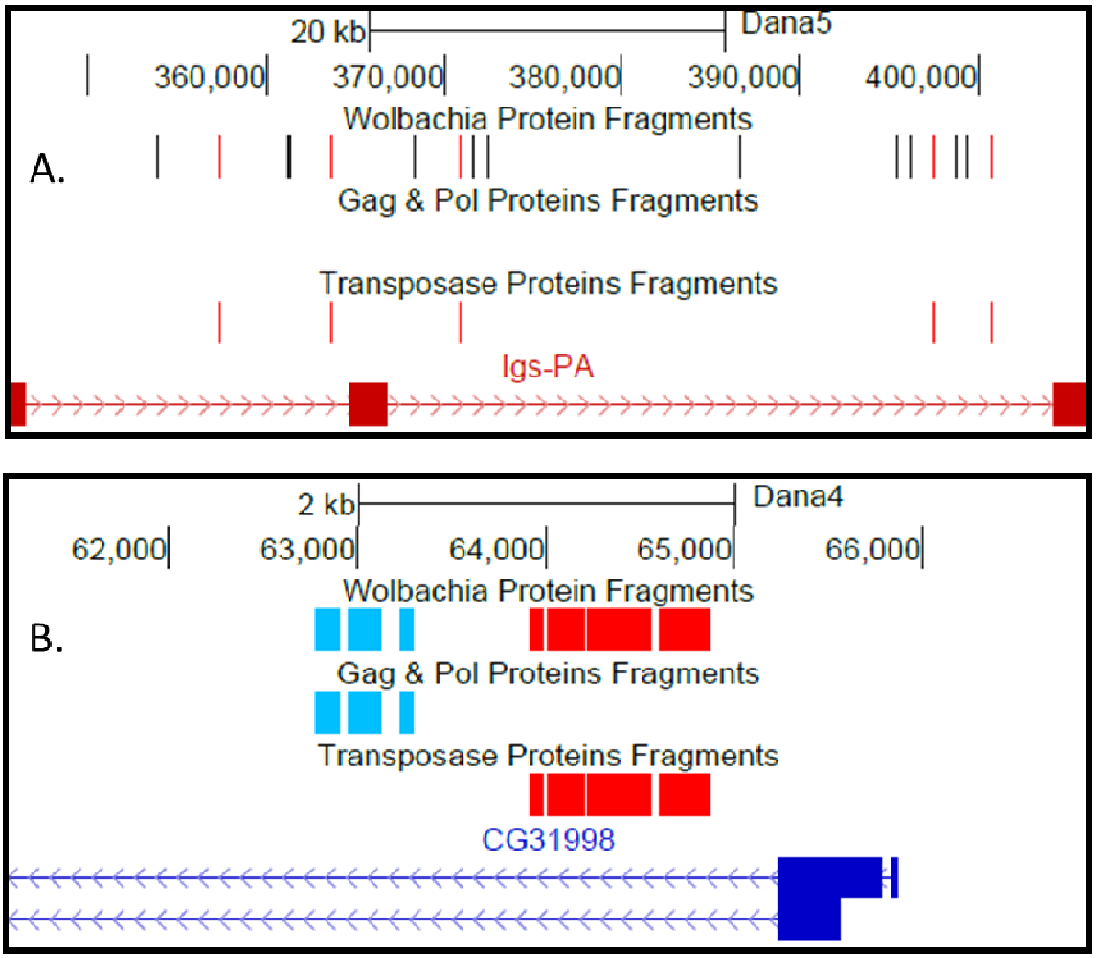


Figure 11: A closer study of the *Wolbachia* protein-coding genes encoding *transposase* in *D. ananassae* F element. A.) The transposase gene fragments are distributed throughout the introns of the expanded *lgs* gene, but all of the transposase gene fragments in *D. ananassae* are located within this one gene on scaffold Improved2_13034. B.) Although less dispersed, all of the *Wolbachia* transposase gene fragments in scaffold Improved_13034 are found in a single gene (*CG31998*).

Transposase overlap with Unknown Repeats

In addition to clustering within the intron region of two expanded *D. ananassae* F element annotated genes, not surprisingly the *Wolbachia* protein-coding genes that encode for transposase also overlap with transposons identified by RepeatMasker. However, while all of the transposase fragments in scaffold improved2_13034 overlap with Unknown repeats, all of the transposase in a different scaffold improved_13034 overlaps with DNA transposons.

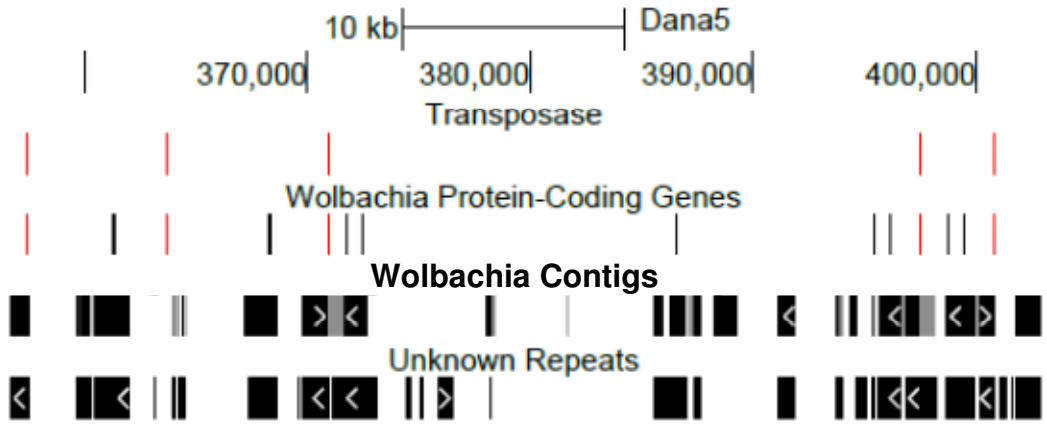


Figure 12: Genome Browser view of *Wolbachia* (*wAna)* contigs and proteins that overlap with Unknown repeats identified by RepeatMasker. 65 out of 117 *Wolbachia* protein-coding genes overlap with Unknown repeats.

In addition to the *Wolbachia* transposase that overlaps with Unknown repeats identified by RepeatMasker, other *Wolbachia* protein-coding gene fragments also overlap with Unknown repeats as well. Of the three scaffolds, 65 *Wolbachia* protein-coding genes out of 117 overlap with Unknown repeats in improved2_13034, None of the *Wolbachia* protein-coding genes out of 10 overlap with Unknown repeats in improved_13034, and 11 of the *Wolbachia* protein-coding genes out of 47 overlap improved_13010 (see Figure 12 for an example). Of these overlaps with Unknown repeats, the *wAna* genes found in each region vary in each of the scaffolds, although gag and pol seem to be frequently present (2 out of 65, 0 out of 0, and 6 out of 11, respectively).

***Wolbachia* overlaps with RepeatMasker**

To see the extent of *Wolbachia* fragments that overlap with transposons identified by RepeatMasker, I calculated the total size of all *Wolbachia* that overlap with RepeatMasker in each of the contigs, as well as the protein-coding genes and their gene fragments (Table 3). The results shows that 25.1% (247068/984715) of the repeats identified by RepeatMasker overlaps with *Wolbachia* contigs. 3.1% (30488/984715) of the repeats identified by RepeatMasker overlaps with *Wolbachia* protein-coding genes.

Scaffold	*Wolbachia* protein-coding genes that overlap with RepeatMasker	*Wolbachia* contigs that overlap with RepeatMasker	Transposons identified by RepeatMasker
Improved2_13034	17,555	98,238	400,854
Improved_13034	3,159	55,688	222,305
Improved_13010	9,774	93,142	361,556
Total	**30,488**	**247,068**	**984,715**

Table 3: Measuring the extent of the total expansion in repeats is due to the *Wolbachia* invasion. The size of each whole scaffold is 467,128bp, 315,470bp, and 597,243bp, respectively.

A more detailed examination of the overlap between the *Wolbachia* contigs and transposons identified by RepeatMasker shows that the *Wolbachia* fragments most frequently overlap with LTR retrotransposons, followed by Unknown repeats, LINEs, and a small number of DNA transposons (Figure 13). Interestingly, only improved2_13034 and improved_13034 had a *Wolbachia* protein-coding gene that aligned with a DNA transposon.

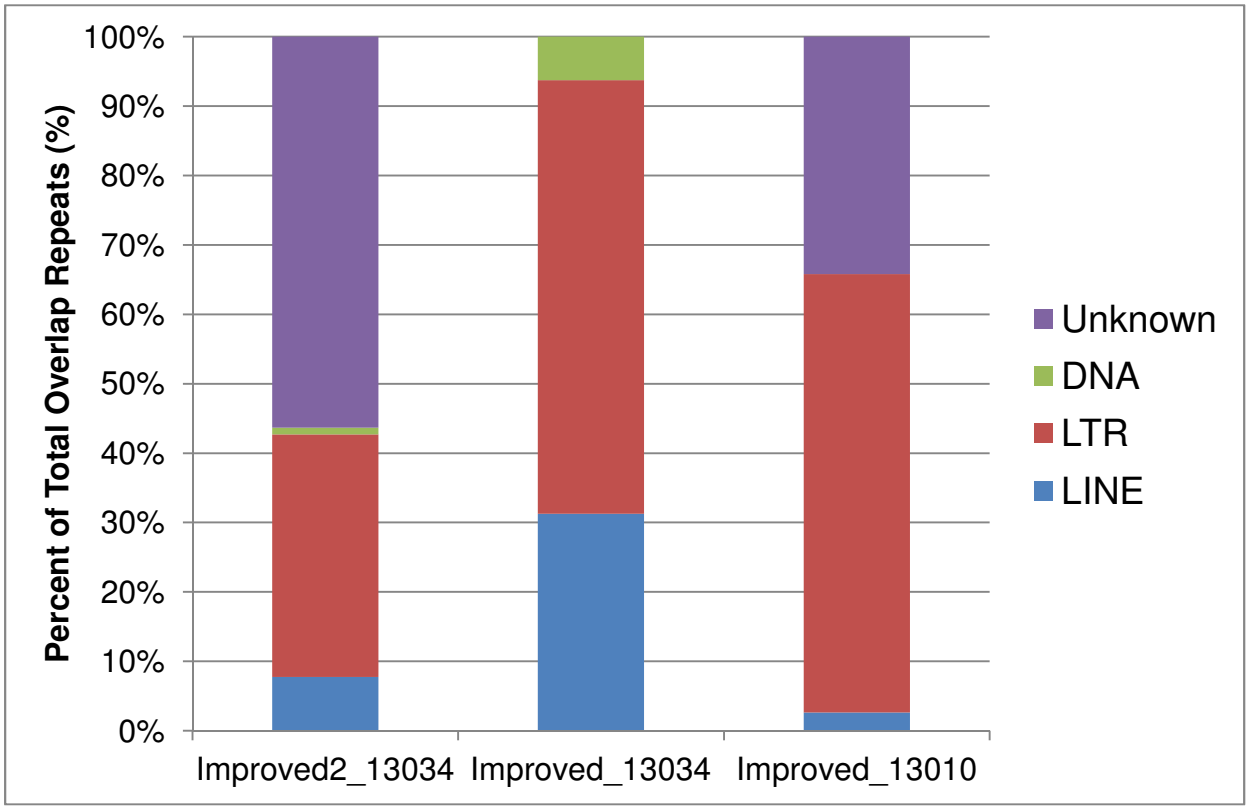


Figure 13: Distribution of *Wolbachia* protein-coding gene fragments which overlap with transposons and other repeats identified by RepeatMasker. The *Wolbachia* protein-coding gene fragments most frequently overlap with LTR (35%, 63%, 63%), Unknown (56%, 34%, 0%), and LINE (8%, 3%, 31%) transposons.

Conclusions

The *D. ananassae* F element is unusual because it has very high repeat density compared to the *D. melanogaster* F element. This high repeat density leads to many misassemblies and gaps in the *D. ananassae* F element assembly. As part of this study, GEP students and I have substantially improved the quality of three F element scaffolds from the *D. ananassae* F element assembly. The sequence improvement process involved resolving misassemblies (e.g. inconsistent mate pairs) and producing additional sequencing data for low quality regions and gaps (Figure 2). For regions that are difficult to sequence, I used multiple PCR techniques to generate PCR products for sequencing (Figure 3). As part of the sequence improvement protocol, GEP students and I used either restriction digests or PacBio reads to confirm the final assemblies, which gives us much stronger confidence in the F element assemblies. Collectively, we were able to assemble and improve ~1.4 Mb of the *D. ananassae* F element. I then used these improved sequences to investigate the expansion of the *D. ananassae* F element.

Because preliminary analysis shows that many regions of the *D. ananassae* F element have strong sequence similarity to *Wolbachia* (*wAna*) contigs and protein-coding genes, I performed a more detailed investigation of the *wAna* genome. My analysis of the *wAna* protein coding-genes shows that *wRi* is a closer informant to *wAna* than *wMel* (Figure 5). This observation was unexpected because the *wAna* was assembled using *wMel* as the reference genome (Salzberg et al, 2005). My results suggest that using *wRi* as the reference genome might improve the overall quality of the *wAna* assembly.

Because my analysis shows that many of the regions in *D. ananassae* F element that show similarity to protein-coding genes in *wAna* are hypothetical proteins, I decided to try to

annotate these *wAna* hypothetical protein-coding genes based on similarity to annotated genes in *wRi* and *wMel*. Using this procedure, I have successfully classified approximately ~30% of the annotated *wAna* hypothetical proteins. From these classifications, I noted that a substantial number (34/133) of the newly annotated *wAna* protein-coding genes were transposase. The findings suggest that *wAna* has a much higher density of transposase genes than other protein-coding genes. In addition, I also find that most of the protein-coding genes in *wAna* are annotated as transposase, gag, and pol proteins.

After investigating the rest of the *wAna* genes, I noticed that the next largest group of *wAna* genes are the IS elements. Examination of the locations of the IS elements in *wAna* shows that they are roughly evenly distributed throughout the entire genome. However, I did observe several contigs in which there were more (i.e. greater than two) IS elements present. Many of these IS elements are clustered together in the *wAna* genome.

My analysis of the *D. ananassae* F element scaffolds did not show any regions with similarity to IS elements, but there are many matches to transposases (which is one of the core components of the IS element). These transposase matches tend to be clustered together either within a single intron (i.e. *CG31998*) or within multiple introns of the same gene (i.e. *lgs*). This difference in the distribution of transposases within the two genes suggests that some *D. ananassae* F element genes might be more susceptible to multiple rounds of lateral transfer of the *wAna* genome than others. F element genes that experienced multiple rounds of lateral gene transfer would likely have larger introns and coding spans.

Further analysis of the *D. ananassae* F element shows that the *wAna* protein-coding genes are more likely to be found in the intronic regions (32% of the intronic regions) than the

intergenic regions (24% of the intergenic regions). This suggests there might be a general preference for *Wolbachia* to transfer into intronic regions. However, due to the small sample size of only 2 genes with transposase and only ~1.4 Mb of the ~20 Mb in *D. ananassae* F element, analysis of additional scaffolds and expanded genes are needed to support this conclusion.

Additionally, my analysis also shows that ~25% of the transposons identified by RepeatMasker overlap with *wAna* contigs on the three *D. ananassae* F element scaffolds. A large portion of the *Wolbachia* overlap with Unknown repeats identified by RepeatMasker, consistent with the hypothesis that *Wolbachia* contributes to the expansion of the *D. ananassae* F element and the lower gene density compared to *D. melanogaster* (Figure 9). However, my analysis also shows that a substantial percentage (35%, 63%, and 63% in improved2_13034, improved_13010, improved_13034) of the *wAna* contigs overlap with LTR retrotransposons identified by RepeatMasker (Figure 13). This could indicate that some of the *wAna* contig might contain a novel class of repeats that has not yet been characterized. Alternatively, this observation could indicate that the *wAna* assembly might be contaminated with LTR transposons that are found in *D. ananassae*. We will need to analyze additional F element scaffolds in order to determine which of the two hypotheses is correct.

In addition to improving the overall *wAna* assembly by using *wRi* as well as analyzing additional F element scaffolds, further work will classify the remaining 194 conserved hypothetical proteins and 259 hypothetical proteins to complete the annotation of the *wAna* assembly. Additionally, because of the recent concern regarding the prior detection of *Wolbachia* DNA in the *D. ananassae* genome (Klasson et al, 2014), we will need to perform polytene squashes and *in situ* hybridization experiments to verify the integration of *Wolbachia* into the *D. ananassae* F element.

Other future experiments include the analysis of *wAna* protein coding genes that are overrepresented in the *D. ananassae* F element to determine how well conserved are they in the other *Wolbachia* species. To do so, I would first measure the rate of evolution by performing a Ka/Ks analysis. Then, in order to identify the conserved regions that might be under selective pressure, I would align all the different copies of the same *Wolbachia* gene on the *D. ananassae* F element against each other using ClustalW2. These conserved regions might correspond to signals that enable *D. ananassae* and *D. melanogaster* F element genes to be expressed within a heterochromatic environment.

We originally became interested in studying the *D. melanogaster* F element because we suspected that it must utilize different mechanisms for regulating gene expression than the other *D. melanogaster* autosomes. Given that the same set of genes are found on both the *D. melanogaster* and *D. ananassae* F elements despite the large difference in repeat density (30% versus 80% repeat), the aberrant signals and mechanisms that allow proper expression of *D. melanogaster* F element genes might be stronger on the *D. ananassae* F element. This hypothesis is supported by the results of a recent study that showed that the coding exons of twelve *D. ananassae* F element genes have very similar properties compared to the orthologous genes in *D. melanogaster* (Thomas Quisenberry, Senior Thesis, WU 2015).

Collectively, our study of the unusual characteristics of the *D. ananassae* F element will improve our understanding of the mechanisms that regulate gene expression in heterochromatic regions. These insights will contribute to our understanding of common human diseases that are caused by the misregulation of gene expression, including cancer (Lee and Young, 2013).

Acknowledgements

Special thanks to Wilson Leung for his assistance and guidance, to Thomas Quisenberry, Kevin Ko, and the GEP students and faculty for their genome contributions, and to the Elgin lab for their support throughout this entire analysis and thesis.

Bibliography

Bartolomé, Carolina, Xulio Maside, and Brian Charlesworth. "On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster." Molecular biology and evolution 19.6 (2002): 926-937.
doi: 10.1093/oxfordjournals.molbev.a004150

Cerveau, Nicolas, et al. "Evolutionary dynamics and genomic impact of prokaryote transposable elements." Evolutionary biology–concepts, biodiversity, macroevolution and genome evolution. Springer Berlin Heidelberg, 2011. 291-312.
doi: 10.1007/978-3-642-20763-1_17

Chandler, M. and Mahillon, J. (2002) Insertion Sequences Revisited: Mobile DNA II. Edited by N.L., Craig et al. ASM Press. Pp 305-366.
doi: 10.1128/9781555817954.ch15

Choi, Jae Young, and Charles F. Aquadro. "The coevolutionary period of wolbachia pipientis infecting drosophila ananassae and its impact on the evolution of the host germline stem cell regulating genes." Molecular biology and evolution (2014): msu204.
doi: 10.1093/molbev/msu204

Clark, Andrew G., et al. "Evolution of genes and genomes on the Drosophila phylogeny." Nature 450.7167 (2007): 203-218.
doi: 10.1038/nature06341

Hotopp, Julie C. Dunning, et al. "Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes." Science 317.5845 (2007): 1753-1756.
doi: 10.1126/science.1142490

Hotopp, Julie C. Dunning, et al. “Response: New names for old strains? Wolbachia wSim is actually wRi” Genome Biol 6.7 (2005):401.
doi: 10.1186/gb-2005-6-7-401

Iturbe-Ormaetxe, Iñaki, Markus Riegler, and Scott L. O'Neill. "New names for old strains? Wolbachia wSim is actually wRi." Genome Biol 6.7 (2005): 401.
doi: 10.1186/gb-2005-6-7-401

Jones, Peter A., and Stephen B. Baylin. "The epigenomics of cancer." Cell 128.4 (2007): 683-692.
doi: 10.1016/j.cell.2007.01.029

Klasson, Lisa, et al. "Extensive duplication of the Wolbachia DNA in chromosome four of Drosophila ananassae." BMC genomics 15.1 (2014): 1097.
doi: 10.1186/1471-2164-15-1097

Klasson, Lisa, et al. "The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans." Proceedings of the National Academy of Sciences 106.14 (2009): 5725-5730.
doi: 10.1073/pnas.0810753106

Lee, Tong Ihn, and Richard A. Young. "Transcriptional regulation and its misregulation in disease." Cell 152.6 (2013): 1237-1251.
doi: 10.1016/j.cell.2013.02.014

Riddle, Nicole C., Christopher D. Shaffer, and Sarah C. R. Elgin. "A Lot about a Little Dot – Lessons Learned from Drosophila Melanogaster Chromosome Four." Biochemistry and cell biology = Biochimie et biologie cellulaire 87.1 (2009): 229–241. PMC.
doi: 10.1139/o08-119

Salzberg, Steven L., et al. "Serendipitous discovery of Wolbachia genomes in multiple Drosophila species." Genome Biol 6.3 (2005): R23.
doi: 10.1186/gb-2005-6-3-r23

Serbus, Laura R., et al "The Genetics and Cell Biology of Wolbachia-Host Interactions." *Annual Review of Genetics* 42 (2008): 683–707.
doi: 10.1146/annurev.genet.41.110306.130354

Siguier, Patricia, et al. "ISfinder: the reference centre for bacterial insertion sequences." Nucleic acids research 34.suppl 1 (2006): D32-D36.
doi: 10.1093/nar/gkj014

Siozios, Stefanos, et al. "Draft genome sequence of the Wolbachia endosymbiont of Drosophila suzukii." Genome announcements 1.1 (2013): e00032-13.
doi: 10.1128/genomeA.00032-13